N° TSE-667

July 2016

"Spatial scale in land use models: application to the Teruti-Lucas survey"

Raja Chakir, Thibault Laurent, Anne Ruiz-Gazen, Christine Thomas-Agnan, and Céline Vignes

Spatial scale in land use models: application to the Teruti-Lucas survey

Raja Chakir[a], Thibault Laurent[b], Anne Ruiz-Gazen[c], Christine Thomas-Agnan[c,*], Céline Vignes[b]

[a] *Economie Publique, INRA, AgroParisTech, Université Paris-Saclay, 78850 Thiverval-Grignon, France*
[b] *Toulouse School of Economics (CNRS), 21 allée de Brienne, 31042 Toulouse, France*
[c] *Toulouse School of Economics, 21 allée de Brienne, 31042 Toulouse, France*

Abstract

We consider the problem of land use prediction at different spatial scales using point level data such as the Teruti-Lucas (T-L hereafter[1]) survey and some explanatory variables. We analyze the components of the prediction error using a synthetic data set constructed from the Teruti-Lucas points in the Midi-Pyrénées region and a five categories land use classification. The study first shows that the number of points in the Teruti-Lucas survey is quite enough for estimating the probabilities of each land use category with a good quality. Furthermore it reveals that, contrary to usual practice, when the objective is to predict land use at aggregated levels, land use probabilities should be estimated at more locations where explanatory variables are available rather than restricting to the initial Teruti-Lucas locations. Indeed this strategy borrows strength from the knowledge of the explanatory variables which may be heterogeneous. Finally, guidelines for constructing the grid of locations for estimation are given from the analysis of the heterogeneity of each explanatory variable.

Keywords: land use models, spatial scale, Teruti-Lucas survey, Gini-Simpson impurity index, classification tree

1. Introduction

It is widely accepted that land use is among the main human pressures on the environment, including greenhouse gas emissions, biodiversity loss and water and soil pollution (Foley & al., 2005). In this context, it is much needed to develop econometric and statistical tools that help to predict the possible land use patterns and trajectories in order to improve our understanding of the causes and consequences of these phenomena.

Land use is determined by complex spatio-temporal interactions between biophysical factors (soil quality, topography) as well as socio-economic factors (population growth, economic conditions and planning). There exist numerous approaches for quantitatively modeling land use patterns in the literature Irwin & Geoghegan (2001); Verburg et al. (2004); Chakir (2015). These approaches can be classified into different groups according to the categories of land use examined (for example rural vs urban, agriculture vs forest, agriculture and urban vs forest), the resolution of the data used (aggregated data vs individual data), the presence or absence of spatial interaction, the inclusion or not of the dynamic dimension, the nature of modeling (statistics, econometrics, geography, cellular automata).

The purpose of most approaches is to provide either insights into the driving factors of land use patterns or prediction of plausible land use patterns. The distinction between these two goals is not always easy and the delimitations of the two approaches are often not easy to distinguish Shmueli (2010).

Difficulties of land use modeling often lie in the frequent lack of data ("good" drivers) and in incompatibility of scale between dependent variable (land use data) and explanatory variables especially for the economic variables (rents, conversion costs and prices). There is a large literature on spatial misalignment and change of support problem (see for example Mugglin et al. (2000); Gotway & Young (2002); Banerjee et al. (2014); Do et al. (2015)). However most of it deals with the prediction of a univariate variable whereas in the land use problem we predict a compositional variable, i.e. a vector of probabilities summing to one.

*Corresponding author
Email addresses: chakir@grignon.inra.fr (Raja Chakir), thibault.laurent@tse-fr.eu (Thibault Laurent), anne.ruiz-gazen@tse-fr.eu (Anne Ruiz-Gazen), christine.thomas@tse-fr.eu (Christine Thomas-Agnan), celine2.vignes@ut-capitole.fr (Céline Vignes)

[1] T-L: Teruti-Lucas

The goal we consider is to define a spatial scale and a strategy for predicting land use at areal level, combining point level data (such as the T-L survey) with a set of explanatory variables which can be at point level or areal level. As we will see in Section 3, one feature of the T-L survey is that the distribution of the sample locations is sparse and does not fill regularly the space. A similar, although not identical, situation happens with the United States Department of Agriculture NRI database Nusser & Goebel (1997) used for example for predicting landscape in Lewis & Plantinga (2007).

Most of the economic variables are collected for administrative units making it simpler to estimate econometric models at the same administrative level. In the case of Teruti-Lucas data, the land use is available at point level for some points distributed on the territory in regular clusters. One knows that the T-L sample is representative of land use at the department level (an administrative division of France, equivalent to NUTS3 regions) and that each observed point represents 100 ha at this aggregated level. One direction of this work is to find an intermediate scale between the points (which are not observed everywhere) and the coarse aggregated departmental level. Indeed, for the assessment of ecological effects of land uses, since the ecological process of interest, such as habitat quality or dispersal of species are relevant at finer scales, a model of land use at the department level within which ecological conditions vary considerably would be less relevant.

Chakir et al. (2016c) propose a simple model explaining the T-L data with easily accessible covariates that allow to predict land use at the point level in five categories. Despite the difficulty of this classification problem, they show that 65% of the locations can be correctly classified using land cover and altitude and that additional variables (meteorological, socio-economic and biophysical, spatially lagged explanatory variables) improve only marginally the prediction. Starting from a good point level model, such as the one obtained in Chakir et al. (2016c), the objective of the present paper is to propose an efficient way for predicting land use at different aggregated levels.

Indeed several strategies are current practice once a good model is fitted to the initial data set. The first is to compute the predictions at point level for the Teruti-Lucas locations and then aggregate these predictions at any larger scale. The second and probably most frequent among practitioners is to average the estimated probabilities at the T-L locations and then derive the predictions from these. But the drawback of these two approaches is to throw away the possible knowledge of explanatory variables outside the T-L locations. We thus imagine and empirically evaluate a third strategy using this knowledge when available: it consists in using the model to estimate the point level probabilities at a fine grid of locations whose density depends on the heterogeneity level of the explanatory variables and then proceed as in the second strategy.

In order to be able to evaluate the efficiency of the different strategies, we define a synthetic data set generated from the Teruti-Lucas survey of 2010. The synthetic data points are 300 meters from each other over the whole Midi-Pyrénées region and the land use, in five categories, is simulated from a model driven by the actual T-L survey locations. The synthetic data allow us to compute prediction errors which would not be possible with a real data set.

Based on this synthetic data, we first analyze the prediction error at the point level by evaluating the proportion of error due to the probability estimation step and the one due to the response prediction step. In our example, it clearly appears that the response error part is dominant while the estimation error is negligible. We also illustrate how the response error is related to the Gini-Simpson impurity index of the point level probability vectors. We then proceed to the analysis of the prediction errors at different levels of aggregation. It reveals that our proposed strategy improves upon the usual practice.

The paper begins in Section 2 by explaining the simulation framework and presenting the different methodological tools useful for analyzing the errors notably the cumulative distribution function of error tolerance and the Gini-Simpson impurity index. In Section 3, we present the Teruti-Lucas survey and a descriptive analysis of our synthetic data set. The results of the analysis of the prediction error are given in Section 4 at the point level and in Section 5 at aggregated levels. Finally Section 6 concludes by the following recommendation for practitioners: when the objective is to predict land use at a given aggregated level from the Teruti-Lucas survey (or alternatively from a model derived from this survey, as the model presented in Chakir et al. (2016c)), one should first look at the spatial scale of explanatory variables to determine a fine grid of locations, then use the model to compute estimated probabilities on this fine grid of locations before averaging them to obtain the predictions at any larger scale.

2. Simulation framework and methodological tools

In order to investigate the questions raised in the introduction, we devise a simulation framework as follows. Starting from an actual data set from the Teruti-Lucas survey (see Section 3), after estimating a classification tree model as in Chakir et al. (2016c,b), we construct a synthetic data set simulating from this model. More precisely, the vector of the land use probabilities estimated by our classification at location i will be used as true parameter in our Data Generating Process of the synthetic data (hereafter DGP). The purpose of creating the synthetic data set is that it can be simulated at any location in space where the explanatory variables are available and we will show later on that this allows to validate strategies different from and better than the classical ones for predicting land use at large scales from point data. From now on, one replication of the synthetic data will play the role of true data. The land use variable at the T-L locations of this synthetic data will be used for fitting the point level model and the land use variable at the remaining locations will be used for validation only.

2.1. Fitting a model

Be it for estimating the initial model from the true data set (in order to build the DGP of the synthetic data) or for fitting a model to the synthetic data set, we use the same technique described below.

The land use probabilities can be estimated by several methods available for categorical variables with more than two categories. In Chakir et al. (2016c,b), we compare multinomial logit models (Train, 2009) and classification trees (Breiman et al., 1984), and get very similar results in terms of percentage of good predictions. We have decided to use classification trees for the present study.

Classification trees are decision trees where the target variable is categorical. The general idea is to use categorical and/or continuous predictors to split the sample into successive nodes corresponding to homogeneous subsets by recursive binary partitioning, so that the set of leaves, or terminal nodes, form a partition of the sample in classes that are expected to be homogeneous. Each node corresponds to one of the predictors chosen by the algorithm. The split is done according to the nature of the predictor: subset of categories for categorical variable, cut-off otherwise. The prediction for an individual of the sample is determined by following the path from the root to the leaf given the values of the predictors for that individual. We use the CART algorithm (Breiman et al., 1984) which first builds a maximal tree and then prune it. We implement it with the R function `rpart` from the **rpart** package (Therneau et al., 2014), with Gini impurity (see Section 2.4.3) as a measure of the quality of the split.

From a classification tree, we can derive estimated probabilities using the empirical frequency corresponding to the group of leaves associated with each land use category. A prediction of land use for each point is usually obtained using the majority rule in each terminal node, which means allocating the most frequent land use to each terminal node.

2.2. Simulating synthetic data from the initial model

The model we get after fitting the classification tree to the initial data, $p_i = f(x_i)$, links the probability vectors of land use with a set x_i of explanatory variables. Even though these probabilities have been estimated at the previous step, we don't use a hat notation since now they represent the true parameters of our DGP for the synthetic data. Note that these can be computed at any location in space where the explanatory variables are available (not only the observed locations of the initial data set) and this is true in the context of our simulation as well as in real applications. We will use a fine grid of points on the territory such that all explanatory variables are known on this grid. We then demonstrate the advantage of using the knowledge of explanatory variables on this fine grid as opposed to restricting attention to the T-L grid points for estimations at aggregated levels. From this set of probabilities, in order to obtain a simulated land use data, we need a prediction method. We will code the land use variable using dummies as follows: $d_{ik} = 1$ if land use k ($k = 1, \ldots, K$) is obtained at location i and $d_{ik} = 0$ otherwise. For reasons which will become clear later, we consider two types of predictions

- "prediction by random draw" using the multinomial distribution with parameters 1 and p_i as above and denoted d^r_{ik} ($d^r_{ik} = 1$ if land use k is chosen by the random draw and 0 otherwise),
- "maximum probability prediction" denoted by d^m_{ik} ($d^m_{ik} = 1$ if p_{ik} is the maximum probability among the p_{ij} $j = 1, \ldots, K$ and 0 otherwise). Maximum probability prediction consists in predicting that land use with maximum probability over all uses.

For the validation process, we generate 1000 replicates of the land use variable d^r_{ik} at all points of the fine grid.

2.3. Fitting a model to the synthetic data

Once a synthetic data set is obtained and a classification tree fitted, because we mimic what would be done from a true data set, we need to fit a model to the synthetic data and also define land use predictions for the locations of our fine grid (once again not only the observed locations of the initial data set). The vector of estimated probabilities is denoted by $\hat{p}_i$ and the predictions which are done as above either by random draw or by maximizing the probability are denoted respectively by $\hat{d}^r_i$ and $\hat{d}^m_i$. Note that the maximum probability prediction is called the Bayes classifier and is known to minimize the Bayes risk (Hastie et al., 2009, p. 21).

2.4. Analyzing the error at point level

2.4.1. Error decomposition

If we assume that the data are generated by random draw (d^r_i) and the predictions are obtained by the maximum probability prediction ($\hat{d}^m_i$), we can look at the vector of errors $\hat{d}^m_i - d^r_i$. It can be decomposed into

$$\hat{d}^m_i - d^r_i = \hat{d}^m_i - \hat{p}_i + \hat{p}_i - p_i + p_i - d^r_i$$

where $\hat{d}^m_i - \hat{p}_i$ is the "estimated response error", $\hat{p}_i - p_i$ the "estimation error" and $d^r_i - p_i$ the "response error".

It is important to notice that the three terms of this error decomposition are of a different nature. The response error and the estimated response error are due to categorization of probabilities into predictions while the estimation error depends on the quality of the model which links the probabilities with the explanatory variables. Thanks to the synthetic data, all the terms of this decomposition can be calculated and compared in terms of sum of squares on our example.

More precisely, the Sum of Squared Errors (SSE) between $\hat{d}^m_{ik}$ and d^r_{ik} defined by $SSE = \sum_{i=1}^{n} \left(\hat{d}^m_{ik} - d^r_{ik}\right)^2$ can be decomposed into:

$$\sum_{i=1}^{n} \left(\hat{d}^m_{ik} - d^r_{ik}\right)^2 = \sum_{i=1}^{n} \left(\hat{d}^m_{ik} - \hat{p}_{ik}\right)^2 + \sum_{i=1}^{n} \left(\hat{p}_{ik} - p_{ik}\right)^2 + \sum_{i=1}^{n} \left(d^r_{ik} - p_{ik}\right)^2 + C \tag{1}$$

where the remainder term C is equal to

$$-2 \sum_{i=1}^{n} \Big[(\hat{d}^m_{ik} - \hat{p}_{ik})(\hat{p}_{ik} - p_{ik}) - (\hat{d}^m_{ik} - \hat{p}_{ik})(d^r_{ik} - p_{ik}) - (\hat{p}_{ik} - p_{ik})(d^r_{ik} - p_{ik}) \Big].$$

The different terms will be given in Table 3 in Section 4.1 for our data example and will be compared.

2.4.2. Cumulative Distribution Function of Error Tolerance (CDFET)

For the purpose of assessing the predictive accuracy of shares in a discrete choice model, Haaf et al. (2014) suggests using the Cumulative Distribution Function of Error Tolerance which is the empirical cumulative distribution function of the absolute value of the response error at point level for each category. This graph explores the complete distribution of the point level error and allows comparisons between the errors for different categories (land use here).

2.4.3. Point level Gini-Simpson impurity index

For a population of individuals classified into K categories, the Gini-Simpson impurity index (Simpson, 1949) is defined by $gs = 1 - \sum_{k=1}^{K} p_k^2$ where p_k is the probability of category k. This index is low (purity) if one probability is very high and all others are low, and it is high (impurity) if all categories have similar probabilities. Note that the index $1 - gs$ is equal to the probability that two individuals taken at random from the data set of interest are of the same category. Both indices are used in many fields for example in ecology for measuring biodiversity, and in economics under the name of Herfindahl index to measure competition. In statistics, it is also used for classification trees under the name of Gini impurity index (not to be confused

with the Gini concentration index). In our case, we want to measure how homogeneous or diverse is land use at a given point or in a given region, with the idea that classification is going to be more difficult when there is diversity.

At point level, we denote this index by $gs_i = 1 - \sum_{k=1}^{K} p_{ik}^2$ for point i where p_{ik} is the probability of land use k at location i. Note that when predictions are obtained using a classification tree, the number of values of gs_i is finite and less than or equal to the number of terminal nodes.

2.5. Analyzing the errors at aggregated levels

In Section 3.2, we define several grids dividing our territory into regular cells (squares actually). We denote by G_g a generic square of such a grid where g indexes the cells. For each cell, we need to define how we aggregate the point level probabilities and predictions to obtain aggregated level probabilities and predictions.

2.5.1. Aggregated probabilities

For each cell G_g, we define three aggregated probabilities:

- $\bar{p}_{gk}$ denotes the average of the probabilities p_{ik} derived from our initial model $p_i = f(x_i)$ for the points i that belong to the same cell G_g: $\bar{p}_{gk} = \frac{1}{\#G_g} \sum_{i \in G_g} p_{ik}$ where $\#G_g$ denotes the number of points in the cell G_g.
- $\bar{p}_{gk}^{dr} = \frac{1}{\#G_g} \sum_{i \in G_g} d_{ik}^r$, where we recall that d_{ik}^r is the prediction by multinomial random draw
- $\bar{p}_{gk}^{dm} = \frac{1}{\#G_g} \sum_{i \in G_g} d_{ik}^m$, where we recall that d_{ik}^m is the maximum probability prediction

Note that $\bar{p}_{gk}$ can be considered as our true aggregated probability whereas $\bar{p}_{gk}^{dr}$ and $\bar{p}_{gk}^{dm}$ are two different empirical estimations of these probabilities. Moreover, in Section 5, we will consider two versions of $\bar{p}_{gk}$ depending on whether we use only the T-L locations or a larger number of locations (those from the fine grid).

2.5.2. Aggregated errors

At aggregated level, we analyze two different kinds of error: the response error and the sampling error. The response error is the error due to categorization of probabilities into predictions analyzed here from an aggregated point of view. The sampling error is the error made using only T-L locations instead of using all the points.

One can think of two different ways of aggregating the response error at the level of a cell G_g. In the first case we look at the difference between true and estimated aggregated probability after aggregation $|\bar{p}_{gk}^{dr} - \bar{p}_{gk}|$ (we will refer to it as to the "aggregated absolute response error"). In the second case we compute the errors at the level of the point and then aggregate them to obtain the "average point level absolute response error" $\frac{1}{\#G_g} \sum_{i \in G_g} |d_{ik}^r - p_{ik}|$.

The sampling errors can only be calculated at aggregated levels because they compare the aggregation of probabilities either of the fine grid points or of the T-L locations. The sampling error is thus defined as $\bar{p}_{gk}(T) - \bar{p}_{gk}^{dr}$ where $\bar{p}_{gk}(T)$ is the aggregation of the probabilities in group G_g at the T-L locations and $\bar{p}_{gk}^{dr}$ the aggregation of the predictions by random draw d_{ik}^r at the fine grid locations.

We analyze these two kinds of errors with RMSE and CDFET as described below. In Section 5, we compute Root Mean Squared Error (RMSE) at different aggregation levels. We will use the generic notation G_g for a given cell of a given aggregation level, where g indexes these cells.

The aggregated probabilities $\bar{p}_{gk}$ are compared to the aggregated probabilities $\bar{p}_{gk}^{dr}$ with RMSE for cell G_g by:

$$RMSE(\bar{p}_{gk}, \bar{p}_{gk}^{dr}) = \sqrt{\frac{1}{NK} \sum_{k=1}^{K} \sum_{g \in I_G} \#G_g \left(\bar{p}_{gk} - \bar{p}_{gk}^{dr}\right)^2},$$

with $\#G_g$ the number of squares in G_g of the baseline grid level, $\bigcup_{g \in I_G} G_g$ is a partition of the region and N the total number of cells G_g. Comparisons of other kinds of aggregated probabilities are done in the same way with RMSE in Section 5.

At aggregated level, the Cumulative Distribution Function of Error Tolerance plots, for each land use, the percentage of squares with an error less than a specified value. This analysis is conducted for both kinds of aggregation for the response error (aggregated response error and average point level absolute response error).

3. Teruti-Lucas data and synthetic data description

Our region of interest is the Midi-Pyrénées region which was (up to December 2015) the largest region in France (8.3% of the whole territory and 3020 municipalities in January 2013) and is made of 8 departments (see Figure 1). It is a quite rural region which accounts for 4.5% of the metropolitan population only (in 2011), but presents the advantage of a diversity in land uses with a major urban center, Toulouse, large farming areas in the middle, the Pyrénées mountains in the South, and pastures and forests in the North.

Figure 1: Midi-Pyrénées region and its 8 departments

3.1. Teruti-Lucas and explanatory variables description

Land use data come from the Teruti-Lucas survey (SSP, n.d.), which is conducted each year since 1982 by the "Service de la Statistique et de la Perspective" of the French Ministry of Agriculture. This survey aims at monitoring evolutions of land use/cover from a representative sample of points across the French territory. The sample was redefined in 2005 to take into account advances in georeferencing of points and to insure compatibility with the LUCAS (Land Use and Coverage Area frame Survey) survey (Eurostat, n.d.) conducted by Eurostat which gathers harmonized data on land use/cover in the European Union. Both surveys combine direct observations made by the surveyors on the ground and satellite images or aerial photographs. Since 2005, the T-L survey contains information about the land use pattern on "segments" containing 10 points used to collect the data each year (see Figure 2). Our sample thus contains 25317 points and 2579 segments for the Midi-Pyrénées region.

The target variable is the land use measured by the Teruti-Lucas "physical occupation" of the land recoded in the following five categories: urban, farming, forests, pastures, natural land. The explanatory variables we consider are coming from diverse data bases at several different scales. Table 1 describes the source of each explanatory variable. More detail about the data sets and variables can be found in Chakir et al. (2016c), in particular the construction of the five categories. The finest information comes from the Corine Land Cover data, and the variables population density and altitude.

Following the initial spacing of 300m between Teruti-Lucas points, we constructed a fine grid, called the complete grid, so that each segment contains 200 locations vs. 10 T-L locations, for a total of 502205 points. The precise construction of this grid is detailed in the following section.

Figure 2: Teruti-Lucas survey: segments and points

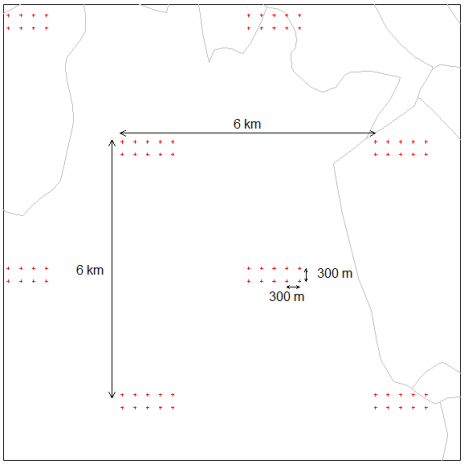


Table 1: Data sources

Name	geographical level	source	year	unit
Land use	6 km segment	Teruti-Lucas	2010	-
CLC2	zones (>25 ha)	Corine Land Cover	2006	-
Altitude	grid (250 m)	BDAlti (IGN)	-	meters
Soil texture	UCS zones	BGSF (GISSOL)	1998	-
Meteorology	grid 25x25km	Agri4cast	2010	-
annual minimum of daily temperature				degrees C
annual maximum of daily temperature				degrees C
annual mean of daily temperature				degrees C
annual sum of rain quantity				millimeters
Land and empty meadow price	32 NRA	Agreste	2010	actual euros/ha
Meadow (more than 70 ha)				
Population density	grid (200 m)	Insee	2010	inhabitants/km^2

3.2. Grids

As we mentioned earlier, we consider regular square grids at several scales, as summarized in Table 2. Level A_0 corresponds to the Teruti-Lucas points, and the coarsest one, denoted by A_7, corresponds to the whole Midi-Pyrénées region. Level A_1 is constructed so that each cell contains a unique T-L segment and so as to tile the territory. Its squares are centered at the barycenter of the 10 points of the T-L segment (see Figure 3) and will be called unit squares hereafter. Their sides have a length of 4.2 kilometers and this grid comprises 2579 such squares. We then construct several successive aggregations of these unit squares until level A_6 where each square contains 1024 unit squares, at each step four squares are gathered together into a single one. Below level A_1 we define three sub-grids, denoted by A_{01}, A_{02} and A_{03}, by dividing each cell of a given level into four squares. For example, an A_1 level square unit is divided into four A_{03} level squares.

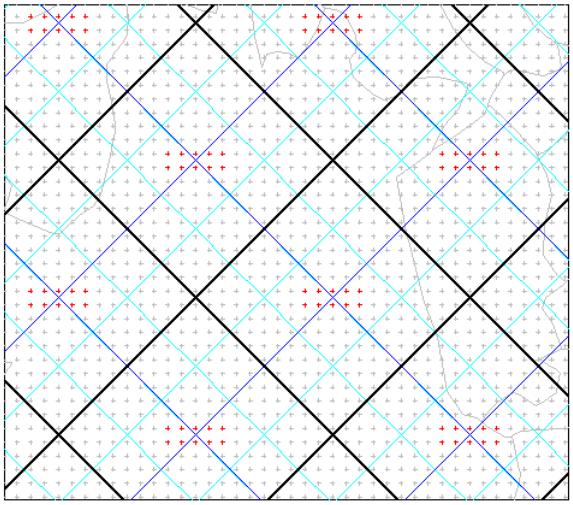
Table 2: Meshes characteristics

Grid	Number of aggregated unit squares	Area (in km^2)	Number of points observed per square	Number of points simulated per square	Total number of squares
A_{01}	1/64	0.2812	0 to 4	1 to 5	161337
A_{02}	1/16	1.125	0 to 4	1 to 13	40608
A_{03}	1/4*	4.5	0 to 4	1 to 50	10246
A_1	1	18	1 to 10	32 to 200	2579
A_2	4	72	1 to 40	32 to 800	689
A_3	16	288	4 to 160	73 to 3200	192
A_4	64	1152	10 to 640	130 to 12800	59
A_5	256	4608	184 to 2559	3528 to 51200	20
A_6	1024	18430	184 to 6605	3528 to 131400	8
A_7	2579	45586.7	25317	502205	1

3.3. DGP for synthetic data

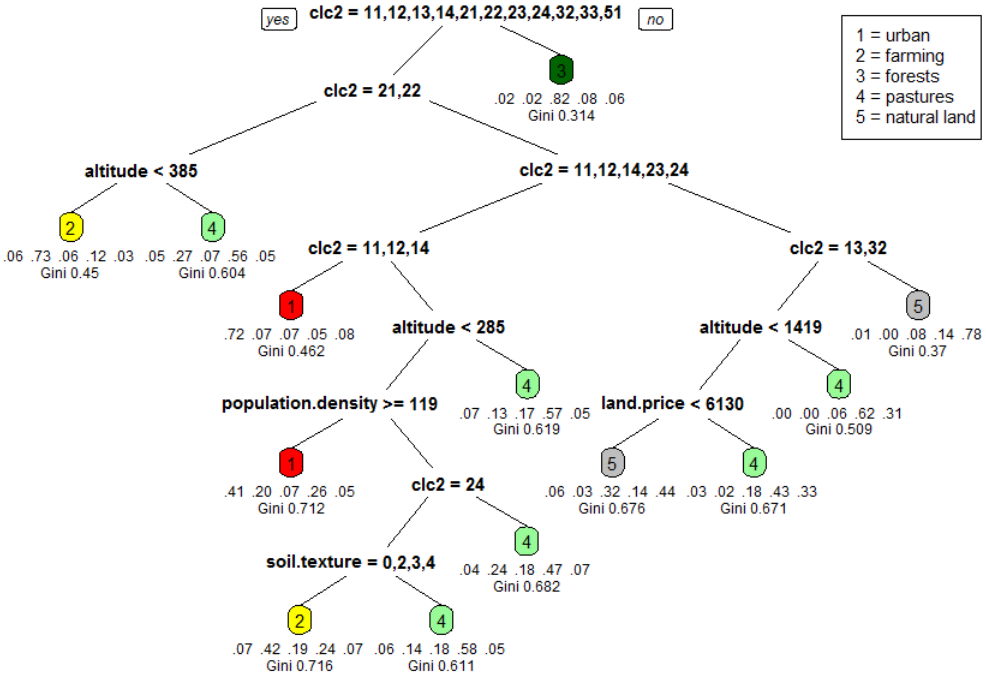
The model chosen for generating the data had to be simple and to reflect some spatial variability with predictors available at different spatial scales. After considering several models from Chakir et al. (2016c,b),

Figure 3: Grids (A_{02} in cyan, A_{03} in blue and A_1 in black) and points (Teruti-Lucas locations in red) in Toulouse area



we have selected the classification tree Tree-E (see Figure 4) and we refer the reader to Chakir et al. (2016c,b) for more details on the description and justification of this model. Motivated by the economics literature on land use, predictors introduced in this classification tree are: *CLC2*, *altitude*, *population density*, *soil texture*, *land price*, *minimum temperature*, *maximum temperature*, *mean temperature* and *rain quantity*. After pruning, the final tree has 13 nodes based on 5 variables: *CLC2*, *altitude*, *population density*, *soil texture* and *land price*.

Figure 4: Classification tree Tree-E, chosen for the DGP



Probabilities p_{ik} are then estimated for all points of the complete grid by the empirical frequency corresponding to the group of leaves associated with each given land use category. Finally, d_{ik}^m are obtained by maximum probability prediction and, for each replicate of the synthetic data, d_{ik}^r are obtained by multinomial random draw with p_{ik} as parameters.

4. Results at point level

4.1. Error decomposition at point level

From the DGP, we have the vectors of probabilities p_i and the vectors of dummies d_i^r at each location i of the complete grid. In what follows, we will consider that the vector of dummies d_i^r is the observed variable to explain and we will compare two situations. In the first situation the dummies are considered as observed only at the Teruti-Lucas locations while in the second situation the dummies are known on the complete grid. The first situation is more realistic since only the T-L points are available in practice. The comparison brings insight on whether the T-L locations are enough to fit the classification tree.

We observe the explanatory variables at each point of the complete grid. We estimate two classification trees using either the complete grid or the T-L locations. The two trees differ very little between them and are also very similar to the DGP tree. Most of the splits and cutoffs are the same. Comparing the two fitted trees, only three cutoff values are slightly different and there is one more split for the T-L tree but the classification results are equivalent. So we can conclude that the T-L points are enough to fit our classification tree. In what follows we do not consider further the tree derived from the whole grid.

From the tree fitted to the T-L points we obtain some estimated vectors of probabilities $\hat{p}_i$ and we can predict the land use at any location of the complete grid by taking the land use associated with the maximum probability $\hat{d}_i^m$. For the following two tables, averages of the errors over the 1000 replicates are reported together with the corresponding standard error below between brackets. The sum of squared errors between the observed land uses and their estimation can be decomposed as in Equation (1) and the different terms are given in Table 3.

Table 3: Decomposition of SSE between $\hat{d}_{ik}^m$ and d_{ik}^r

	urban		farming		forests		pastures		natural land	
	mean (SE)	%	mean (SE)	%	mean (SE)	%	mean (SE)	%	mean (SE)	%
$\sum_i (\hat{d}_{ik}^m - \hat{p}_{ik})^2$	5859.3 (348.5)	17.6	27287.9 (725.9)	32.2	12224.1 (451.4)	17.1	37196.2 (875.9)	33.4	8783 (558.5)	21.0
$\sum_i (\hat{p}_{ik} - p_{ik})^2$	28.8 (27.9)	0.1	57.1 (33.2)	0.1	83.5 (87.4)	0.1	117.2 (71.5)	0.1	68.8 (56.6)	0.2
$\sum_i (d_{ik}^r - p_{ik})^2$	27462.1 (137.7)	82.3	57461.3 (131.2)	67.7	59265.3 (162.7)	82.8	74044.4 (141.5)	66.4	32862 (135.4)	78.5
remainder	14.4 (307.8)	0.0	14.6 (702.4)	0.0	-6.6 (414.1)	0.0	123.8 (793.1)	0.1	141.7 (392.8)	0.3
$\sum_i (\hat{d}_{ik}^m - d_{ik}^r)^2$	33364.7 (197.9)	100.0	84820.9 (286.9)	100.0	71566.4 (460.6)	100.0	111481.6 (450)	100.0	41855.4 (501.6)	100.0

On average, the response error $d_{ik}^r - p_{ik}$ and the estimated response error $\hat{d}_{ik}^m - \hat{p}_{ik}$ appear to be the largest terms, whereas the estimation error $\hat{p}_{ik} - p_{ik}$ is negligible with respect to SSE. The average remainder term is also very small in comparison with the two response errors. It is not surprising to get a very small sum of squares for the estimation error since the data have been generated using a classification tree with the same variables as the ones used to fit the model. The standard errors are reasonably small for the large contributions, except for the remainder term but this does not invalidate the fact that the response error is clearly the dominant one.

Finally, the response error is larger than the estimated response error. This last point was expected since the prediction using the maximum probability is known to minimize the SSE and hence its error should be better than that of the random draw prediction. So, the response errors are large and deserve particular attention. In order to simplify the rest of the analysis, we decide not to consider the estimation error any longer and only focus on the response errors $d_i^m - d_i^r$ and their decomposition without taking into account the estimation stage. In practice, with the true T-L data set, there is no way to evaluate the estimation error and we can just conjecture that the response error is also more important than the estimation one. Table 4 gives the decomposition of this response error where we see that the main contribution is due to the discrepancy between the probability and the random draw.

4.2. Analysis of the response error at point level

Since the response error is the dominant one, we further analyse it by looking at the distribution of its absolute value $|d_{ik}^r - p_{ik}|$ for each land use and by relating it to the Gini-Simpson index.

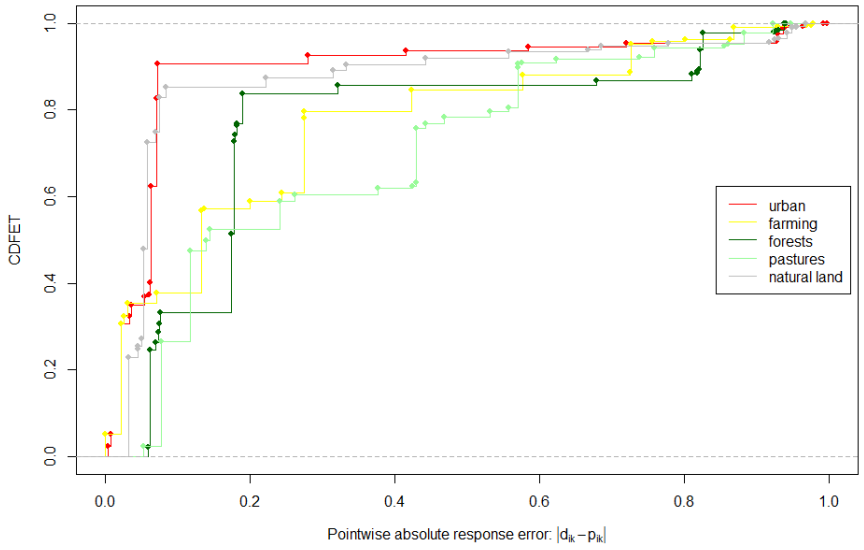
Table 4: Decomposition of SSE between d^m_{ik} and d^r_{ik}

	urban		farming		forests		pastures		natural land	
	mean (SE)	%	mean (SE)	%	mean (SE)	%	mean (SE)	%	mean (SE)	%
$\sum_i (d^m_{ik} - p_{ik})^2$	5919.5 (0.0)	17.7	27307.6 (0.0)	32.2	12138.5 (0.0)	17.0	37327.7 (0.0)	33.5	8867.3 (0.0)	21.2
$\sum_i (p_{ik} - d^r_{ik})^2$	27463.7 (140.8)	82.3	57462.0 (134.3)	67.8	59270.1 (158.1)	83.0	74047.1 (134.1)	66.5	32862.4 (139.5)	78.7
remainder	1.6 (67.7)	0.0	-2.2 (156.7)	0.0	0.8 (82.8)	0.0	9.0 (184)	0.0	8.2 (87.8)	0.0
$\sum_i (d^m_{ik} - d^r_{ik})^2$	33384.9 (173.4)	100.0	84767.5 (242.8)	100.0	71409.5 (233.2)	100.0	111383.8 (264.3)	100.0	41737.9 (185.6)	100.0

4.2.1. CDFET for absolute response error

Figure 5 shows the CDFET which plots, for each land use, the percentage of points with absolute response error less than a specified threshold. For low values of the error threshold, the curves are very similar and we observe errors for all categories less then 0.1 for approximately 35% of the points. These good predictions correspond to homogeneous zones with either low or high probability. This trend remains the same for increasing error threshold for the case of urban and natural land since medium probabilities are quite rare and we have 90% of errors less than 0.1. For farming, forests and pastures, the presence of medium probabilities causes a deterioration of the curve behavior. The long flat part of the forests curve between 0.1 and 0.2 can be explained by the gap in the distribution of p_{i3} as we can see on Figure 6.

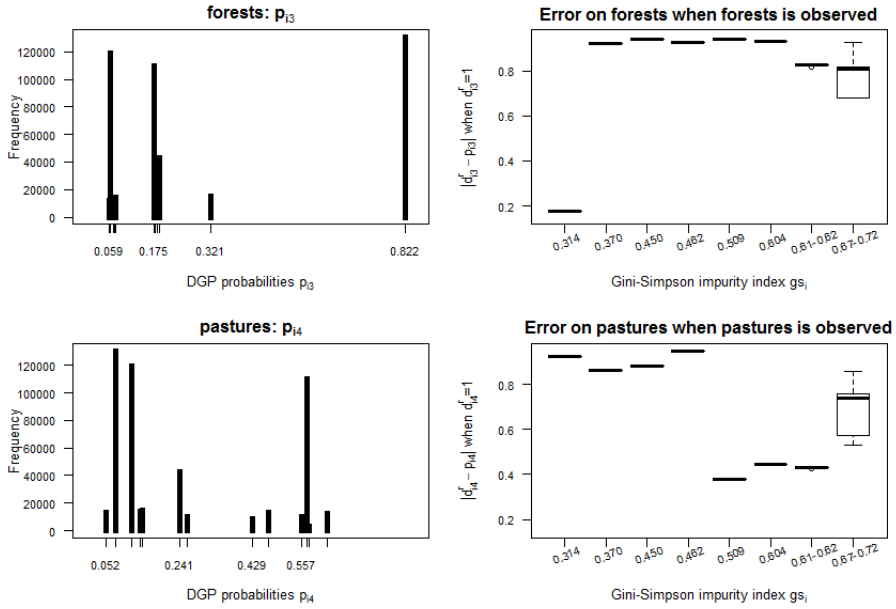
Figure 5: CDFET for pointwise absolute response error ($|d^r_{ik} - p_{ik}|$)



4.2.2. Relationship between the absolute response error and the Gini-Simpson index

We suspect that it is going to be more difficult to predict at locations where land use is heterogeneous, and hence that we should observe a relationship between the errors and the impurity measured by the Gini-Simpson impurity index at point level. Due to the nature of the model (classification tree), the DGP probabilities p_{ik} are discrete with values corresponding to each terminal node of the tree of the DGP. This implies that the values of the Gini-Simpson impurity index (see subsection 2.4.3) are also discrete. Note that low values of the index correspond to purity and homogeneous groups in terms of land uses. In the remainder of the present subsection and in order to be concise, we focus on the forests and pastures land uses but more detail can be found in Chakir et al. (2016a). On its left panel, Figure 6 gives the DGP probabilities for the forests (at the top) and pastures (at the bottom) land uses. The probabilities are always quite low for pastures compared to forests where there exists a large number of points with a large probability (0.822). At such points the impurity index is low (0.314) which means homogeneity in terms of land use. On the right panel of Figure 6, parallel boxplots are displayed for the forests land use at the top and the pastures land use at the bottom. These

Figure 6: Left panel: DGP probabilities, right panel: Absolute response error vs the Gini-Simpson index (for forests at the top and pastures at the bottom).



boxplots give the absolute response error for forests (resp. pastures) use for locations where the forest (resp. pastures) use is observed as a function of the Gini-Simpson index. For the forests we note that there is one boxplot showing low errors while the remaining ones display large errors. The low error boxplot corresponds to the value 0.314 of the Gini-Simpson index with points in a rather pure environment, where the main land use is forests and hence a value of p_{i3} close to 1 which explains the low error. On the contrary, for the other values of the Gini-Simpson index which correspond to points in a non purely forests environment, the error is large because the probability p_{i3} is low. For the pastures land use, the situation differs because they are situated in more heterogeneous (impure) regions. We simultaneously see errors which are neither very low nor very large. Similar conclusions can be drawn for the other land uses as detailed in Chakir et al. (2016a). In particular, the urban land use behaves similarly to the forests one.

To summarize, the prediction error at point level is essentially due to the response error in this synthetic data set with different patterns across land uses. Moreover, the analysis reveals that the larger errors correspond to heterogeneous zones.

5. Results at aggregated levels

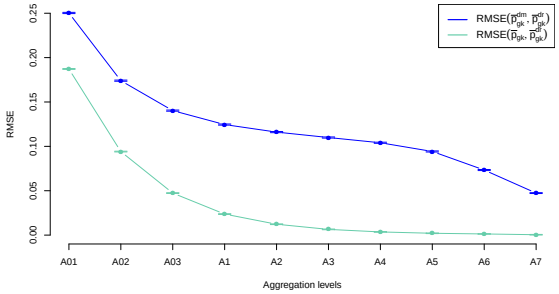
We now turn attention to analyzing the errors at aggregated levels. As mentioned in Section 4.1, the synthetic data we consider is d_i^r at the point level for the complete grid (see Section 2.5.1) and the aggregated data for a given level of aggregation with cells G_g is $\bar{p}_{gk}^{dr}$. We are going to analyze the aggregated response error in the same way as for the point level in order to compare them. Moreover we study the sampling error (as defined in Section 2.5.1) which appears only at aggregated levels.

5.1. Response error

Figure 7 presents the evolution across the different aggregation levels of

- the average RMSE($\bar{p}_{gk}^{dm}$,$\bar{p}_{gk}^{dr}$) over the 1000 replicates, which expresses the distance between the aggregation $\bar{p}_{gk}^{dm}$ in group G_g of maximum probability predictions d_{ik}^m and the aggregation $\bar{p}_{gk}^{dr}$ in group G_g of predictions by random draw d_{ik}^r,
- the average RMSE($\bar{p}_{gk}$,$\bar{p}_{gk}^{dr}$) over the 1000 replicates, which expresses the distance between aggregated probabilities $\bar{p}_{gk}$ in group G_g from point level probabilities p_{ik} and the aggregation $\bar{p}_{gk}^{dr}$ in group G_g of predictions by random draw d_{ik}^r.

Figure 7: Comparison of RMSE($\bar{p}_{gk}^{dm}$,$\bar{p}_{gk}^{dr}$) and RMSE($\bar{p}_{gk}$,$\bar{p}_{gk}^{dr}$)



The average RMSE($\bar{p}_{gk}$,$\bar{p}_{gk}^{dr}$) is lower than the average RMSE($\bar{p}_{gk}^{dm}$,$\bar{p}_{gk}^{dr}$) at all levels and their dispersion is negligible compared to their difference. Hence it is clearly better to directly aggregate estimated probabilities rather than aggregate point level predictions. For levels A_2 to A_7 the average RMSE($\bar{p}_{gk}$,$\bar{p}_{gk}^{dr}$) is very good but it increases at levels A_1 and A_{03} and gets rather poor at lower levels. This last point can be explained by the low number of locations at lower levels.

5.2. Comparison of aggregated response error and average point level response error with CDFET

In Figure 8, we consider aggregations at the A_1 level of the response error in two different ways. In the first case we consider the difference between aggregated probabilities at the A_1 level $|\bar{p}_{gk}^{dr} - \bar{p}_{gk}|$ and call it the "aggregated absolute response error". In the second case we consider the average point level absolute response error $\frac{1}{\#G_g}\sum_{i\in G_g}|d_{ik}^r - p_{ik}|$ at the A_1 level.

Figure 8: CDFET for aggregated absolute response error at A_1 level ($|\bar{p}_{gk}^{dr} - \bar{p}_{gk}|$) on the left, and CDFET for average point level absolute response error ($\frac{1}{\#G_g}\sum_{i\in G_g}|d_{ik}^r - p_{ik}|$) at A_1 level on the right

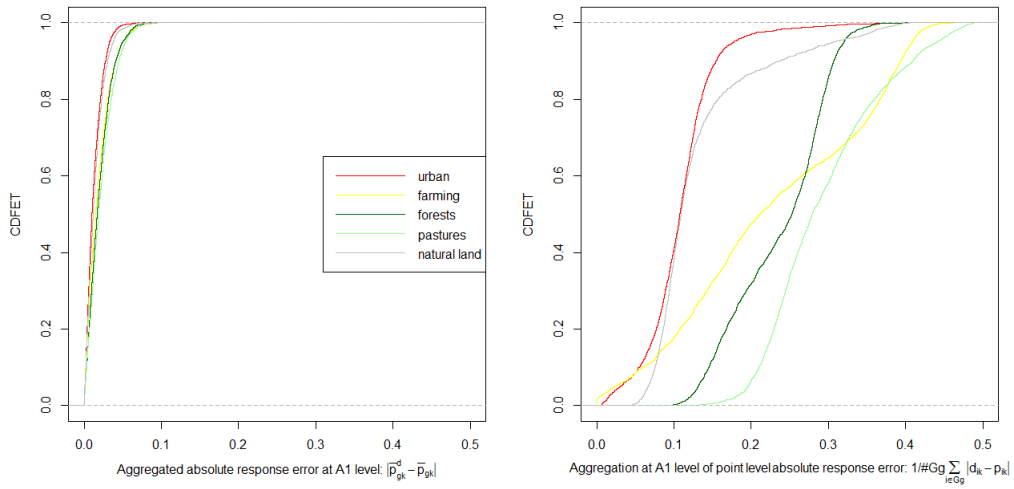
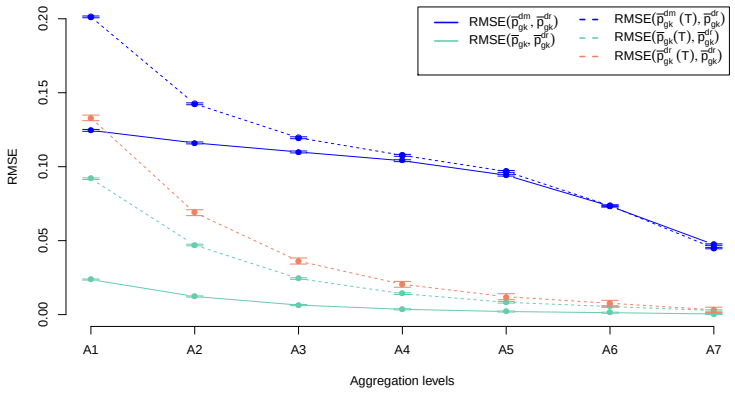


Figure 8 shows that the aggregated response error (left panel) is much lower than the average point level response error (right panel). Results on the aggregated response error are very good for all land uses. Even if all curves are very similar, results are slightly better for urban and natural land: all segments have an error less than 5% for urban use and natural land and less than 10% for other land uses. When considering the average point level response error (right panel), results are worse, even for the urban use and natural land: overall 80% of the segments have an aggregated error less than 13% for urban and 15% for natural land, less than 30% for forests and 35% for pastures and farming.

5.3. Sampling error

In this section, we will compare the aggregated probabilities in a given group G_g based on Teruti-Lucas points data, $\bar{p}_{gk}(T)$, to the aggregated probabilities $\bar{p}^{dr}_{gk}$ which are computed from all locations of our fine grid. The error $\bar{p}_{gk}(T) - \bar{p}^{dr}_{gk}$ can be considered as a "sampling error" and we analyze it across the different spatial levels.

Figure 9: RMSE comparison of the aggregation of maximum probability predictions at all points ($\bar{p}^{dm}_{gk}$) and at Teruti-Lucas locations ($\bar{p}^{dm}_{gk}(T)$), aggregated probabilities at all points ($\bar{p}_{gk}$) and at Teruti-Lucas locations ($\bar{p}_{gk}(T)$), and the aggregation of predictions by random draw at Teruti-Lucas locations ($\bar{p}^{dr}_{gk}(T)$) to the aggregation of predictions by random draw at all points ($\bar{p}^{dr}_{gk}$)



In Figure 9, the dashed green curve plots the evolution across the aggregation levels of the average $RMSE\left(\bar{p}_{gk}(T), \bar{p}^{dr}_{gk}\right)$ which is the RMSE between the aggregated probabilities in group G_g for the T-L points ($\bar{p}_{gk}(T)$) and the aggregation of the predictions by random draw at all locations ($\bar{p}^{dr}_{gk}$). Its dispersion is very small compared to the distances between average curves. This average RMSE corresponds to the sampling error, it is quite important at A_1 level but it is halved at each of the next two aggregations levels and shows a moderate decrease for the following aggregation levels. The dashed green and blue curves show that the corresponding RMSE are higher when the probabilities are aggregated using the T-L locations only (dashed curves) than when they are aggregated using all locations (solid curves, already seen in Figure 7), even if the differences decrease when the aggregation level increases. The interpretation is the same for the dashed salmon curve which represents the average $RMSE\left(\bar{p}^{dr}_{gk}(T), \bar{p}^{dr}_{gk}\right)$ and is quite high compared to zero for the first aggregation levels. From level A_3, the differences between aggregations using T-L points and aggregations using all the fine grid points decline for any aggregation method (aggregation of estimated probabilities in green or aggregation of maximum probabilities predictions in blue). At large scales (A_4 or A_5), it is equivalent to use the T-L points only or the fine grid but we want to emphasize that at lower scales the benefit of using the fine grid is large especially at level A_1. Concerning the T-L points, the curves first confirm the fact that it is better to aggregate estimated point level probabilities rather than to predict at point level. Moreover, it is clear that if a point prediction is used before aggregation, it is better to choose the random draw prediction rather than the maximum probability. We note that the distance between the blue and the green curves is the same for solid (fine grid) and dashed curves (T-L grid), which means that the discretization has the same effect for the fine grid or the T-L grid. Finally, we observe that in general the RMSE are high at levels A_1 and A_2 for the dashed curves because for these levels the number of points per grid cell is lower than 10 and 40 respectively.

5.4. Sampling error for explanatory variables

The heterogeneity between the results we get using the Teruti-Lucas locations and that using all locations is explained by the high resolution of the explanatory variables and the fact that we can observe them on the fine grid. For a given cell, if a difference between these aggregated probabilities does exist, it is due to the

fact that the explanatory variables take values at the fine grid points which on average on the square do not coincide with the average obtained from the T-L locations alone due to their heterogeneity. Therefore we can borrow strength from this heterogeneity to improve our estimations. We are going to analyze the sampling error for the three most important explanatory variables: land cover $CLC2$ which is a categorical variable, altitude and population density which are continuous variables.

For a categorical variable we define, as for the land use variable, the RMSE sampling error by the difference between the mean of the dichotomized categorical variable over all points of the fine grid and the mean of the dichotomized categorical variable over the T-L points. In Figure 10, the RMSE sampling error of the categorical predictor $CLC2$ is divided by more than two at the first aggregation step and is almost null from the A_3 level.

Figure 10: RMSE for $CLC2$ sampling error

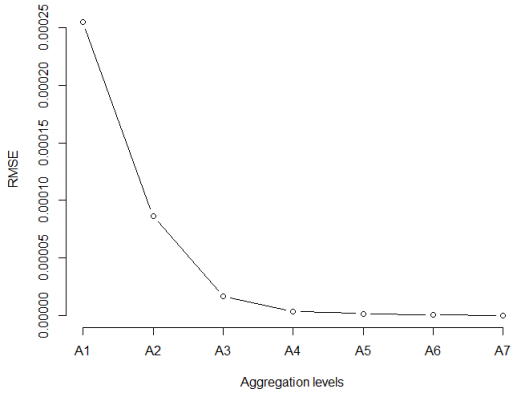
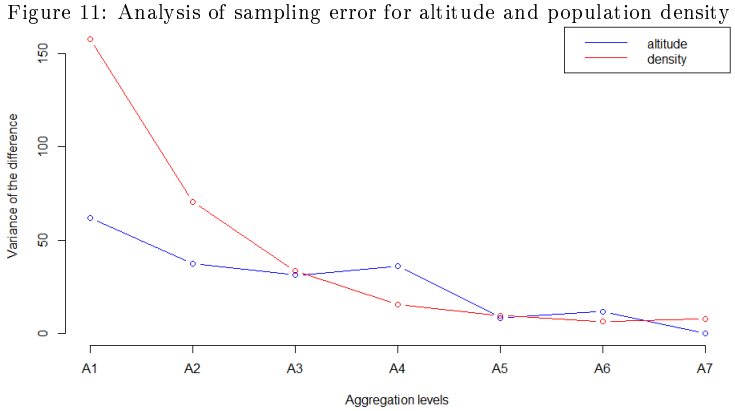


Figure 11 shows the variance of the difference in means between the points of the fine grid and the T-L points for two quantitative variables with high resolution, altitude and population density. For the first two aggregation levels, the sampling error decreases with aggregation, even if the evolution is less strong for the altitude. Except for the A_4 level (where altitude presents a peak due to an artifact of collapsing squares with low altitude and squares with very high altitude in the South of the region), values of both variables for higher spatial levels are low and close to each other.

Figure 11: Analysis of sampling error for altitude and population density



The last two figures can be related to Figure 9. The larger differences in RMSE sampling error are obtained for the A_1 level and A_2 to a lesser extent. From level A_3, the curves are more similar for land use as well as for the three explanatory variables. Hence the analysis of the explanatory variables contains information about the aggregation level at which we can consider that the T-L points are sufficient.

6. Conclusion

At the point level, our synthetic example study illustrates that, even if the land use probabilities are very well estimated, the prediction is poor for all the points where none of the land use probabilities are high. This fact shows that predicting the land use in five categories at the point level is a difficult classification problem. So, we advise the data-analyst to work at aggregated levels or regions rather that at point level. It is also recommended to average the point level estimated probabilities without making any land use prediction for points. As soon as the number of points included in the aggregated level is large enough, the strategy of averaging estimated probabilities leads to very good results at least for our example.

Another interesting conclusion from our paper in the context of land use is that the Teruti-Lucas survey contains enough observation points in order to fit a good classification model for estimating the vector of probabilities. However, averaging the estimated probabilities obtained at the T-L locations only, as we did in Chakir et al. (2016c), is not very efficient. Since the explanatory variables we use in our model are available at any location in the Midi-Pyrénées region and not only at the T-L locations, one is better off using this information. The importance of estimating the probabilities at many locations is clearly illustrated by our results. It is particularly essential at aggregated levels where the explanatory variables exhibit some heterogeneity. Thus the analysis of the heterogeneity of the explanatory variables should help in defining the grid points on which the land use probabilities have to be estimated.

In the future, we plan to compare this approach based on aggregating point level probability estimation with some models fitted directly at the aggregated level such as regression models for compositional data.

Acknowledgments. This work was partially financed by the Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005). We thank the Service de la Statistique et de la Prospective of the French Ministry of Agriculture for letting us use the Teruti-Lucas survey data in the framework of the ANR ModULand, the INFOSOL laboratory of INRA for the pedological data (BDGSF) and the Joint Research Center-MARS of the European Commission for the meteorological data (Interpolated Meteorological Data Base).

References

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data.* (2nd ed.). CRC Press.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Chakir, R. (2015). L'espace dans les modèles économétriques d'utilisation des sols: enjeux méthodologiques et applications empiriques. *Revue d'Économie Régionale & Urbaine*, *1*, 59–82.

Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., & Vignes, C. (2016a). Exploring land use prediction errors with the gini-simpson impurity index. *Case Studies in Business, Industry and Government*, . Submitted.

Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., & Vignes, C. (2016b). Land use predictions on a regular grid at different scales and with easily accessible covariates. URL: `https://www6.versailles-grignon.inra.fr/economie_publique/Media/fichiers/Working-Papers/Working-papers-2016/WP_16_01` working Paper.

Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., & Vignes, C. (2016c). Prédiction de l'usage des sols sur un zonage régulier à différentes résolutions et à partir de covariables facilement accessibles. *Revue Economique*, . To appear.

Do, V. H., Thomas-Agnan, C., & Vanhems, A. (2015). Spatial reallocation of areal data - another look at basic methods. *Revue d'économie régionale et urbaine*, *1/2*, 27–58.

Eurostat (n.d.). Land cover/use statistics (lucas) overview. `http://ec.europa.eu/eurostat/web/lucas/overview`. Accessed: 2016-06-09.

Foley, J. A., & al. (2005). Global consequences of land use. *Science*, *309*, 570–574.

Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, *97*, 632–648.

Haaf, C. G., Michalek, J. J., Morrow, W. R., & Liu, Y. (2014). Sensitivity of vehicle market share predictions to discrete choice model specification. *Journal of Mechanical Design*, *136*, 121402.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer.

Irwin, E. G., & Geoghegan, J. (2001). Theory, data, methods: developing spatially explicit economic models of land use change. *Agriculture, Ecosystems & Environment*, *85*, 7–24.

Lewis, D. J., & Plantinga, A. J. (2007). Policies for habitat fragmentation: combining econometrics with gis-based landscape simulations. *Land Economics*, *83*, 109–127.

Mugglin, A. S., Carlin, B. P., & Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, *95*, 877–887.

Nusser, S. M., & Goebel, J. J. (1997). The national resources inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, *4*, 181–204.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, (pp. 289–310).

Simpson, E. (1949). Measurement of diversity. *Nature*, *163*, 688.

SSP (n.d.). Teruti-lucas, utilisation du territoire. `http://agreste.agriculture.gouv.fr/enquetes/territoire-prix-des-terres/teruti-lucas-utilisation-du/`. Accessed: 2016-06-09.

Therneau, T., Atkinson, B., & Ripley, B. (2014). *rpart: Recursive partitioning and regression trees*. URL: `http://CRAN.R-project.org/package=rpart` r package version 4.1-8.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge Books. Cambridge University Press.

Verburg, P. H., Schot, P. l., Dijst, M. J., & Veldkamp, A. (2004). Land use change modelling: current practice and research priorities. *GeoJournal*, *61*, 309–324. URL: `http://dx.doi.org/10.1007/s10708-004-4946-y`. doi:`10.1007/s10708-004-4946-y`.